# Todd_Garner_DS6306_Week6_FLS

Todd Garner

2023-02-05

For Live Session Week 6 Part 1 - **Download the training set: Connect to the opendatasoft website and download the random sample of 891 Titanic Passengers. This is the training set. The data come in JSON form format and you can use this URL to access the data:**

https://public.opendatasoft.com/api/records/1.0/search/
?dataset=titanic-passengers&rows=2000&facet=survived&facet=
pclass&facet=sex&facet=age&facet=embarked

Hint: This is not trivial. I recommend that you use the jsonlite package (fromJSON()) and RCurl package (getURL()) to access the data. (We covered this in Unit 4).

1. *Try your best to access the data using the URL. You may also find the data (titanic_train.csv) on github. We will go over this data ingestion in live session.*

```
## Warning: package 'jsonlite' was built under R version 4.
```

# Load the Titanic dataset

data(“Titanic”)

# Change the values of the “Survived” column to “Lived” or “Died”

$$\text{Titanic}Survived = ifelse(Titanic\text{Survived == 1, “Lived”, “Died”})$$

# Create a new dataset with only the “Age” and “Class” columns

```
Titanic_subset = Titanic[, c(“Age”, “Class”)]
```

# Impute missing values in the “Age” column with the median

Titanic_subset$Age = ifelse(is.na(Titanic_{s}ubset$Age),
median(Titanic_subset$Age, na.rm = TRUE), Titanic_{s}ubset$Age)

# Split the data into a training set and a test set

```
set.seed(123) train_index = sample(1:nrow(Titanic_subset),
floor(0.8 * nrow(Titanic_subset))) train =
Titanic_subset[train_index, ] test = Titanic_subset[-train_index, ]
```

# Fit a K-NN model on the training set

```
library(class) knn_model = knn(train[, -2], test[, -2], train[, 2], k = 3)
```

# Predict the survival of passengers in the test set

pred = knn_model

# Evaluate the accuracy of the model

mean(pred == test[, 2])

I will have to modify the code to load the training data set.

*Many hours later after trying to mold the ChatGPT code into a workable solution, discarding it and following an article on Edureka.co, I've thrown up my hands. The Edureka explanation makes perfect sense but I'm just not able to get it to work. So, I'm back here at the start to contemplate, "What is it that I'm really trying to accomplish?" I've used the Kaggle data sets and I've defaulted back to the .csv files in the GitHub repo. So, I'm on #2.*

**Use KNN to classify those who survived and died based on Age and class.** In other words, find a way to model whether a passenger lived or died based on Age and Pclass (1st, 2nd or 3rd class passage). Solving for lived or died based on these two columns and the values within them. I will search my code to see where I've gone wrong.

```
##   PassengerId Survived Pclass
## 1           1        0      3
```